DOCUMENT RESUME

ED 408 344 TM 026 626

AUTHOR	Tirri, Henry; And Others
TITLE	Using Neural Networks for Descriptive Statistical Analysis of Educational Data.
PUB DATE	Mar 97
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Classification; *Data Analysis; *Discriminant Analysis; *Educational Research; Foreign Countries; Mathematical Models; Probability
IDENTIFIERS	*Mixture Density Networks; *Neural Networks

ABSTRACT

Methodological issues of using a class of neural networks called Mixture Density Networks (MDN) for discriminant analysis are discussed. MDN models have the advantage of having a rigorous probabilistic interpretation, and they have proven to be a viable alternative as a classification procedure in discrete domains. Both classification and interpretive aspects of discriminant analysis are discussed, and the approach is compared to the traditional method of linear discriminants as implemented in standard statistical packages. It is shown that the MDN approach performs well for both aspects. Many of the observations made are not restricted to the particular cases at hand, and are applicable to most applications of discriminant analysis in educational research. (Contains 31 references.) (Author/SLD)

A paper presented at the 1997 Annual Meeting of the American Educational Research Association (Chicago, IL, USA, March 1997)

Using Neural Networks for Descriptive Statistical Analysis of Educational Data

Henry Tirri and Tomi Silander
Complex Systems Computation Group (CoSCo)*
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland

Kirsi Tirri
P.O. Box 38,Department of Teacher Education
FIN-00014 University of Helsinki, Finland

Abstract

In this paper we discuss the methodological issues of using a class of neural networks called Mixture Density Networks (MDN) for discriminant analysis. MDN models have the advantage of having a rigorous probabilistic interpretation, and they have proven to be a viable alternative as a classification procedure in discrete domains. We will address both the classification and interpretive aspects of discriminant analysis, and compare the approach to the traditional method of linear discriminants as implemented in standard statistical packages. We show that the MDN approach adopted performs well in both aspects. Many of the observations made are not restricted to the particular case at hand, and are applicable to most applications of discriminant analysis in educational research.

*URL: http://www.cs.Helsinki.FI/research/cosco/

1 Introduction

Artificial neural networks (Haykin, 1994) can be viewed as a family of nonlinear models used for empirical regression and classification modeling. Such networks have been successfully used in various fields for nonlinear modeling and approximation, for example in speech recognition (Kohonen, 1995), expert systems (Gallant, 1993), and machine vision(Hinton and Sejnowski, 1983). More recently they have also been applied for data analysis in various financial domains (Baestaens et al., 1994). The increasing importance of neural networks as nonlinear models is witnessed by the fact that currently many of the standard statistical software packages include feed-forward neural network modeling in their tool box. Similarly the recent multidisciplinary research efforts in the field of "Knowledge Discovery in Databases" (Fayyad et al., 1996) use quite frequently neural network techniques.

Neural network models are composed of a large number of individual computational elements called *nodes*, which are linked together to form a structure (called the *architecture*). This structure typically classifies the neural network type: feed-forward neural networks are layered structures, whereas recurrent neural networks introduce feedback links down the network. The nodes are associated with a nonlinear function $y = f(x)$, and the links have associated *weights* $\vec{w}$. The computation is organized by combining the weights with inputs, i.e., multiplying the input value x_i by the corresponding value w_i, which is then given as argument for f. Thus the computation for a single node is given by

$$y = f(\sum_{x_i} w_i x_i).$$

Intuitively, the weights are the parameters of the model, and the *learning* of a neural network from a data sample means parameter estimation. The description above is a gross simplification of this very rich set of models, but captures the essential idea.

Giving an introduction to the various different types of neural networks and their related learning algorithms (parameter estimation methods) is outside the scope of this paper, and an interested reader should consult one of the excellent text books available (Bishop, 1995; Haykin, 1994; Ripley, 1996), or many of the introductions to neural networks from a statistical perspective (see e.g., (Cheng and Titterington, 1994; Ripley, 1993)). Two reviews of Hinton (Hinton, 1989; Hinton, 1992) are also valuable. We would like to point out that neural network model analysis can be based on various different viewpoints from particle physics and statistical modeling to biological simulation, or automata theory. *Our approach to neural networks is based on seeing them as probabilistic models*, which, as opposed to some other views, gives us a rigorous underlying theoretical framework for their analysis and use. In this sense we conform to the work presented in (Bishop, 1995; Mackay, 1992; MacKay, 1992a; MacKay, 1992b; Jordan and Jacobs, 1994).

In spite of their widespread use for data modeling in economics, physics, computer science and pattern recognition, neural networks are almost unknown in the educational sciences community. This is perhaps partly caused by the unfamiliar terminology associated with neural networks due to their origin from cognitive and neurosciences, partly by the lack of demonstrations of the applicability of the methods for educational data. Since many neural network models assume continuous input variables (outputs), educational data sets, such as questionnaire data, have not been modeled due to their discrete nature. In this paper we focus on the use of a particular class of neural networks called the Mixture Density Networks (Bishop, 1994) in the analysis of educational data. This intuitively very appealing neural network model family is introduced in Section 2, and can be understood as an implementation of a particular subclass of finite mixture models (Titterington et al., 1985).

Klecka (Klecka, 1981) defines discriminant analysis to be a set of statistical techniques to study the differences between two or more groups of objects with respect to several variables simultaneously. In educational research discriminant analysis is used for two different purposes:

- **Interpretation of group differences**—i.e., to find out if one is able to discriminate between the groups on the basis of some set of characteristics. In addition one might be interested in finding which characteristics are the most powerful discriminators.
- **Classification**—i.e., predicting the group membership of new data for which the group information is not known.

In Section 5 we will discuss the use of Mixture Density Networks for the classification problem formulated in Section 3. Instead of just presenting classification accuracy information, we want to put the results in perspective, and compare them to the ones achieved by the traditional linear discriminant analysis (McLachlan, 1992). We would like to point out that the purpose here is not to demonstrate the superiority of the Mixture Density Network approach in classification accuracy (although, due the power of the underlying mixture model language, this in many cases is the case). Rather we would like to discuss methodological issues for both constructing the classifiers and for evaluating their quality, if one moves from the linear discriminant framework to neural network approaches. Many of the concerns raised are well-known in the computational intelligence community (Bezdek, 1994), but seem to be very seldom discussed in the educational quantitative methodology literature.

Finally we will address the interpretive side of the discriminant analysis. Since any model that predicts well has captured an underlying regularity in the data, an interesting question is whether that information can be extracted from the model representation. Traditional neural networks suffer from the fact that the language of weight matrix and node functions is not easily interpretable, and results in a "black box" approach, which clearly is not useful in most cases for educational data analysis. In Section 6 we will briefly illustrate that this is not the case for Mixture Density Networks (due to their probabilistic semantics), and discuss the interesting explorative possibilities offered by the MDN models.

We aim at keeping the technical level of our discussion at as moderate level as possible, and focus on discussing the methodological issues using a typical example data sets, one of them being from a recent educational study. Readers not interested in the technical details of the MDN network models can browse Sections 2 and 3, and go directly to the description of the problem domains (Section 4), from which the data samples for the experiments were taken.

2 Mixture Density Networks

Mixture Density Networks (Bishop, 1994) is a neural network class which can be used to represent general conditional probability densities $p(\vec{t}|\vec{d})$ by considering a (semi)parametric model for the distribution of $\vec{t}$, whose parameters are determined by the outputs of a feed-forward neural network, which takes $\vec{d}$ as its input. Thus the MDN models are actually a combined neural network structure and a density model (for more details see the discussion in (Bishop, 1995)). Provided we consider a sufficiently flexible network, and a sufficiently general density model, we have a framework for approximating arbitrary conditional distributions.

Typical choices for a parametric model are a single Gaussian or a linear combination of fixed set of kernel functions. A very general framework for modeling unconditional distributions can be based on the set of *discrete finite mixtures* ((Everitt and Hand, 1981), (Titterington et al., 1985)), where the joint domain probability distribution is approximated as a weighted sum of mixture distributions.

Let $X_1, \ldots, X_m$ be a set of m ($m \geq 1$) discrete (random) variables, and $\vec{d} \in D$ is a sample from the joint distribution of the variables $X_1, \ldots, X_m$. Then the *finite mixture* distribution for $\vec{d}$ can be written as ($K \geq 1$)

$$\begin{aligned} p(\vec{d}) &= p(X_1 = x_1, \ldots, X_m = x_m) \\ &= \sum_{k=1}^{K} \left(p(Y = y_k) p(X_1 = x_1, \ldots, X_m = x_m | Y = y_k) \right), \qquad (1) \end{aligned}$$

where Y denotes a latent *clustering random variable*, the values of which are not given in the data D, and K is the number of possible values of Y.

Thus in finite mixture models the problem domain probability distribution is approximated by a weighted sum of mixture distributions, where each mixture component $p(X_1 = x_1, \ldots, X_m = x_m | Y = y_k)$ models one data producing mechanism. If the variables $X_1, \ldots, X_m$ are independent, given the value of the clustering variable Y, equation (1) becomes

$$p(\vec{d}) = \sum_{k=1}^{K} \left(p(Y = y_k) \prod_{i=1}^{m} p(X_i = x_i | Y = y_k) \right). \qquad (2)$$

For the Mixture Density Networks considered here this independence assumption holds and consequently computation uses equation (2).

A finite mixture model partitions the data to K clusters. This partitioning can be modeled by introducing for each data vector $\vec{d_j}$ an unobserved latent variable Z_j, the value of which gives the the cluster index for the cluster vector $\vec{d_j}$ belongs to. We can now think a vector $Z = (z_1, \ldots, z_N)$, consisting of the values of the latent variables $Z_1, \ldots, Z_N$, as a random sample from the distribution of Y like D is a random sample from the joint distribution of $X_1, \ldots, X_m$. However, for technical reasons it is more convenient to consider each value z_j as a vector of *cluster indicator variable* values, $z_j = (z_{j1}, \ldots, z_{jK})$, where

$$z_{jk} = \begin{cases} 1, & \text{if } \vec{d_j} \text{ is sampled from } P(\cdot | Y = y_k), \\ 0, & \text{otherwise.} \end{cases}$$

Finite mixtures as defined in equation (2) is a generic model family, as we still have to fix the cluster distribution $p(Y)$ and the intra-class conditional distributions $p(X_i|Y = y_k)$[1]. Most commonly used component functions in the literature are the univariate normal distributions (see e.g., (Titterington et al., 1985)). In educational domains the variables are usually discrete, thus we can drop the assumption of the form of the distribution. For the univariate case a binomial model could be used, but for the general case with $m > 1$ a natural choice is the multivariate generalization of the binomial distribution called the *multinomial distribution*

$$p(\vec{c}|\Theta) = \left(\begin{array}{c} N' \\ c_1 \ldots c_{n_i} \end{array} \right) \prod_{j=1}^{n_i} \theta_j^{c_j}$$

where $\vec{c} = (c_1, \ldots, c_{n_i})$ is the vector of counts of the number of observations of each value of X_i. In addition the sum of probabilities $\sum_{j=1}^{n_i} \theta_j = 1$ and $\sum_{j=1}^{n_i} c_j = N'$ (N' is the total number of observations). Since we are interested in the data distribution, i.e., $p(X_i|Y = y_k)$ the multinomial distribution form simply reduces to a product of probabilities θ_j. Analogously we assume that the cluster distribution $p(Y)$ is multinomial. Thus in order to get a model, we need to fix the number of the mixing distributions (K), and determine the values of the model parameters. For technical reasons it will be convenient to make a notational distinction between the mixture weight parameters and the parameters of the intra-class conditional distributions, i.e., $\Theta = (\alpha, \Phi), \Theta \in \Omega$, where $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\Phi = (\Phi_{11}, \ldots, \Phi_{1m}, \ldots, \Phi_{K1}, \ldots, \Phi_{Km})$, with the denotations $\alpha_k = P(Y = y_k)$, $\Phi_{ki} = (\phi_{ki1}, \ldots, \phi_{kin_i})$, where $\phi_{kil} = P(X_i = x_{il}|Y = y_k)$.

Since our estimation of the network parameters will be Bayesian (Bernardo and Smith, 1994) we need to fix the prior distributions for the parameters. The family of Dirichlet (multivariate Beta) densities is conjugate to the family of multinomials, therefore we assume that prior distributions of the parameters are $(\alpha_1, \ldots, \alpha_K) \sim \mathrm{Di}\,(\mu_1, \ldots, \mu_K)$ and $(\phi_{ki1}, \ldots, \phi_{kin_i}) \sim \mathrm{Di}\,(\sigma_{ki1}, \ldots, \sigma_{kin_i})$, $(1 \leq k \leq K, 1 \leq i \leq m)$, where

$$\{\mu_k, \sigma_{kil} \mid 1 \leq k \leq K; 1 \leq i \leq m; 1 \leq l \leq n_i\}$$

are called the *hyper parameters* of the corresponding distributions. Assuming

[1] Here we consider only mixtures in which all the component distributions come from the same parametric class.

that the parameter vectors α and Φ_{ki} are independent, the joint prior distribution of all the parameters can be expressed as

$$\mathrm{Di}\left(\mu_1,\ldots,\mu_K\right)\prod_{k=1}^{K}\prod_{i=1}^{m}\mathrm{Di}\left(\sigma_{ki1},\ldots,\sigma_{kin_i}\right).$$

The finite mixture model family is universal in the sense that it can approximate any distribution arbitrarily close as long as a sufficient number of components is used. Unfortunately such generality typically implies also that parameter estimation can become computationally inefficient. Therefore the networks used in our experiments will be a special subclass of the general Mixture Density Networks. This class follows from equation (2) when we remove the latency of Y and assume that one of the variables $X_1,\ldots,X_m$ gives us the partitioning of the data (for notational simplicity we will assume that it is always X_m). These new models correspond to a specific subclass of the more general case, thus the joint probability distribution for a data vector $\vec{d}$ can be written as

$$\begin{aligned} p(\vec{d}) &= p(X_1 = x_1,\ldots,X_m = x_m, X_m = k) \\ &= \sum_{j=1}^{n_m}\left(P(X_m = j)\prod_{i=1}^{m-1}p(X_i = x_i|X_m = k)\right). \end{aligned} \qquad (3)$$

3 Classification problem

Let us now return to the classification problem. The purpose of a classification procedure is to predict the value of a single class variable of a new partially observed data vector, based on the model (e.g., discriminant functions) constructed from the sample.

Given the data sample D, MDN predictions are based on the conditional distribution $p(\vec{d}|D)$ of a new *test vector* $\vec{d}$, where

$$p(\vec{d}|D) = \frac{p(\vec{d},D)}{p(D)}. \qquad (4)$$

The classification problem can now be restated: Given the values of the variables $X_1,\ldots,X_{m-1}$, and a data sample D, predict the value of variable X_m. For notational simplicity, in the sequel we drop the variable names, and denote a value assignment $(X_1 = x_1, X_2 = x_2,\ldots,X_{m-1} = x_{m-1})$ by writing

$(x_1, x_2, \ldots, x_{m-1})$. Now for each possible value x_{mi}, $x_{mi} \in \{x_{m1}, \ldots, x_{mn_m}\}$ we wish to compute the probabilities

$$p(X_m = x_{mi} | (x_1, \ldots, x_{m-1}), D).$$

From the Bayes' theorem (Bernardo and Smith, 1994) we know that

$$\begin{aligned} p(X_m = x_{mi} \quad | & \quad (x_1, \ldots, x_{m-1}), D) \\ & = \frac{p(\vec{d}[x_{mi}]|D)}{\sum_{k=1}^{n_m} p(\vec{d}[x_{mk}]|D)}, \end{aligned} \tag{5}$$

where $\vec{d}[x_{mi}]$ denotes the vector $(X_1 = x_1, \ldots, X_{m-1} = x_{m-1}, X_m = x_{mi})$. Consequently, the conditional distribution for variable X_m can be computed by using the complete data vector conditional distributions (4) for each of the possible complete vectors $\vec{d}[x_{mi}]$. The resulting distribution is called the *predictive distribution* of X_m.

The derivation of different possibilities as the predictive distribution $p(\cdot)$ in the case of MDN is somewhat involved and omitted here. The derivations can be found e.g., in (Heckerman et al., 1995; Kontkanen et al., 1997; Tirri et al., 1996). For the present purposes it is enough to state that for the restricted case of finite mixtures discussed in the previous section, the calculation of the predictive distribution can be performed efficiently without any approximations.

4 Data description

For our experiments we used three data sets, one from medical domain (Primary Tumor), one from chemical analysis (Glass) and an educational data set from a recent study (Effectiveness). The Primary Tumor data sets concerns predicting the location of primary tumor, where the location of the cancer is the group variable. Glass Identification database (USA Forensic Science Service) is concerned of grouping glass defined by their oxide content (i.e. Na, Fe, K, etc). Both of these data sets are standard benchmarks for comparing different classification procedures. Since the educational data used is particular to the study at hand, we will give a more detailed description of it.

The educational data used in this study was gathered for the research project "Effectiveness of Teacher Education in Finland" in the spring 1996. The objective of the project was to evaluate the effectiveness of Finnish teacher

Data set	Size	#Variables	#Classes (Groups)
Glass	214	10	6
Primary tumor	339	18	21
Effectiveness	204	42	4

Table 1: The description of the data sets used in our experiments.

education at various levels from individual to international teacher education policy. A more detailed description of the framework and research conducted in the project is discussed in (Niemi and Tirri, 1996). The data adopted to this study was gathered to investigate how well teacher education had been able to achieve the goals set to it. These goals were selected from school-law, programs of teacher education and other documents describing teachers' work at school. The teachers and their educators from four different teacher education departments in Finland were asked to perform self-evaluation on the success of teacher education for helping teachers to achieve these goals. The evaluation instrument consisted of 41 behavior statements (and information about the teacher education department), and used a Likert scale from 1 to 5 for the assertions. The results of this evaluation study are reported in the forthcoming study (Niemi and Tirri, 1997).

The data sample used for our comparison is derived from the teachers' data in the study described above. This data consists of ratings of 204 Finnish teachers. The subjects were teaching at two levels, one half being elementary school class teachers (N=110) and the other half secondary school subject teachers (N=94). These teachers came from four different teacher education departments in three different counties of Finland. The gender distribution was representative to that of Finnish teacher population—25% were males.

A short description of the data sets used can be found in Table 1.[2]

[2] The Primary Tumor and Glass data sets can be obtained from the UCI data repository at URL address "http://www.ics.uci.edu/~mlearn/".

5 MDN in classification

Let us now first study the problem of developing a classification procedure, which would allow us predict the group to which a given data vector most likely belongs. For example for the Effectiveness data set this means developing a model, which would allow us to predict from which of the four different teacher education departments a teacher comes from based on his/her answers to the questions. In the application domain this information is interesting for finding the topics that could be improved in each of the teacher education departments. Here we will allow the classification procedures to use all the 41 predictor variables in constructing the predictive model, which is atypical to a questionnaire data analysis. In practice for this type of problems discriminant analysis is preceded by dimensionality reduction procedures, e.g., factor analysis, and one would use summarized information such as the factor scores instead of the primary variables. Knowing the difficult issues related to selecting a proper factor structure, this would, however, introduce another parameter to our study, i.e., the discriminative quality of the factor variables constructed. Although the analysis is performed at the primary variable level, all discussion is naturally valid for discriminant analysis with factor scores also.

5.1 Testing with sample vs. cross validation

The traditional classification procedures in linear discriminant analysis typically use either the discriminating variables or the canonical discriminant functions constructed from the data (Klecka, 1981). We assume that the reader is familiar with the standard approach as implemented in the SPSS statistical software package (Norušis, 1990), and do not repeat the principles here.

What we are more interested in is the validation of the classification procedure constructed, either by the Linear Discriminant (LD) or by the Mixture Density network approach. For MDN we have described the classification procedure as the calculation of the predictive distribution $p(\vec{d}[x_{mi}]|D)$, which in our case depends on parameters Θ. The corresponding notion to our model Θ in the linear discriminant analysis are the canonical discriminant functions (let us denote them by $\vec{f}_{\mathrm{ld}}$). A typical procedure to test the accuracy of $\vec{f}_{\mathrm{ld}}$ is to classify the cases in the sample from which the model was constructed. We will call this approach *training sample validation.* The resulting percentage of correct predictions together with analysis of the difference to the expected number of correct predictions is then used to quantify the quality of the model.

Although many textbooks include a warning about the fact that testing a model with the same data sample from which it is constructed (see e.g., (Klecka, 1981), pp. 51-52) gives inflated estimates of the classification performance, this seems to be the standard practice, unless the size of the data sample is large and sometimes independent samples are used. In particular, use of k-fold cross validation (Stone, 1974), sometimes known as the "jackknife", tends to be very rare. This is quite concerning, as it is well known that most parameter learning procedures have a tendency to *overfit*, i.e., form classification functions that are more accurate for the sample than they would be for the full population. In particular we will demonstrate that the classification accuracy of both LD and MDN is substantially different for the training sample, than if measured by cross validation. The more parameters the model used in the discriminant analysis has, the more severe this overfitting phenomenon is. With the exception of the simple model class of perceptrons (Haykin, 1994), all neural network model families are highly parameterized nonlinear function estimators, and would perform extremely poorly, if the models were selected based on their training sample performance. Therefore in the computational intelligence community the training sample based validation has been totally replaced by other methods such as cross validation based estimation.

An interesting question is, why has this not happened in the educational research community? The answer is intuitively simple, but has important consequences for the common practices, if neural networks models (or actually any highly parameterized or nonparametric model class) are to be used. The number of parameters for the hyperplanes used in $\vec{f}_{\mathrm{ld}}$ for low-dimensional data spaces is so low that the model is not able to overfit much, and thus automatically shows some generalization to the full population. This can be clearly seen from Table 5.1, where the discriminant function model $\vec{f}_{\mathrm{ld}}$ with the variable selection (5 variables) is only able to fit the model to the training sample to reach 51% accuracy with 45.5% performance in leave-one-out cross validation. Notice that in the 41 variable case we see the difference of 22% for LD between the classification in the training sample and cross validation. Naturally MDN shows the same behavior, although for the more general (semiparametric) MDN the results would be even more illustrative; for this particular data set we could reach over 90% accuracy with the training sample with very poor generalizability when tested with out of sample data.

In Table 5.1 we report the classification accuracy of both the LD and MDN

DATA SET	METHOD	LD (SPSS)	MDN
Effectiveness	**nvs and mdo**		
	training sample	72.0	74.5
	5-fold crossvalidation	46.0	43.0
	leave-one-out crossvalidation	50.0	38.5
	vs and mdo		
	training sample	49.5	58.0
	5-fold crossvalidation	45.0	44.5
	leave-one-out crossvalidation	44.0	45.0
	nvs and mdi		
	training sample	69.0	75.0
	5-fold crossvalidation	44.5	39.5
	leave-one-out crossvalidation	48.5	39.5
	vs and mdi		
	training sample	51.0	57.0
	5-fold crossvalidation	44.0	45.5
	leave-one-out crossvalidation	45.5	45.0
Primary Tumor	**nvs and mdi**		
	training sample	48.0	56.9
	leave-one-out crossvalidation	36.0	49.0
Glass	**nvs and mdi**		
	training sample	64.5	79.0
	leave-one-out crossvalidation	60.3	70.1

Table 2: The comparison of the classification performance of the linear discriminant functions and Mixture Density Networks. The option "nvs" and "vs" denote that no variable selection/variable selection was used, i.e., 41 predictor variables/5 predictor variables were used. Options "mdo" and "mdi" correspond to omitting data with missing values and including missing value as a value in the analysis, respectively. The numbers represent the percentage of correctly classified cases.

methods in the case of the data sets described in Section 4. For comparative purposes, in the Effectiveness case, we have also included the results for a reduced variable set, the selection of the variables was performed by the standard stepwise selection procedure. In addition it is interesting to observe that as opposed to LD, MDN performance improves when the number of variables is decreased. This is due to the fact that instead of pure discrimination, MDN in fact tries to model the full joint distribution of the variables and thus has to balance the predictions for all the variables, not just the group variable.

From the above discussion we would like to stress that *reporting the classification performance in the training sample is in most cases quite misleading*, and definitely not to be used with more complex model families such as neural networks.

5.2 Classification performance vs. training sample size

In the previous Section we saw that for the Effectiveness data set LD outperforms the MDN in the cross validated error rate when all the 41 predictor variables are used, and for that for the 5 predictor variable case both methods showed equal performance. On the other hand for the Primary Tumor and Glass data sets MDN clearly outperforms the standard LD. Let us now study what happens to the performance of these two methods as a function of training sample size.

In this type of experiments one randomly partitions sample in a *training sample reservoir* D_r containing 70% of the sample, and a *test set* D_q containing the remaining 30 %. One data $\vec{d_1}$ is then randomly taken out of the training reservoir and used as a training sample $D_1 = \{\vec{d_1}\}$. This initial training sample D_1 is used to construct the model which is used to classify all $\vec{d_j} \in D_q$, and the predictions thus obtained for each $\vec{d}$ are then compared to the actual outcomes k.

Next the training set D_1 is extended by another data instantiation $\vec{d_2}$, unequal to the element already in D_1, but otherwise randomly picked from the training reservoir D_r. This new training set is denoted by D_2. After building the new model, all $\vec{d_j} \in D_q$ are classified again, and the results are compared tot he actual outcomes. This procedure of adding one training element to D_i to form D_{i+1}, determining the model using D_{i+1} and predicting the value of the group variable for each entry in the test set is then repeated until $D_{i+1} = D_r$, i.e., contain the full reservoir. This whole procedure is then repeated 10 times

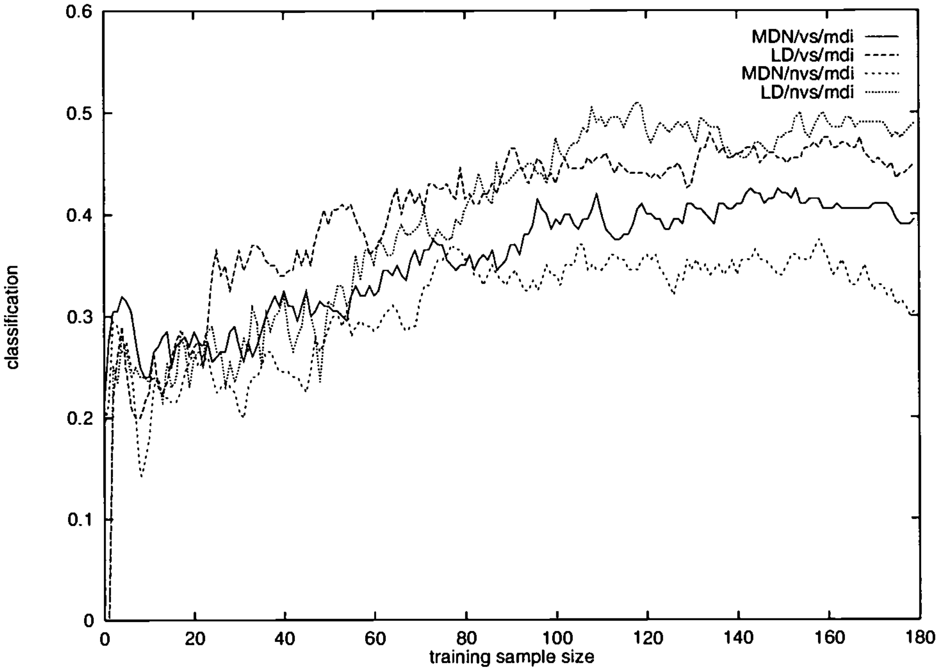
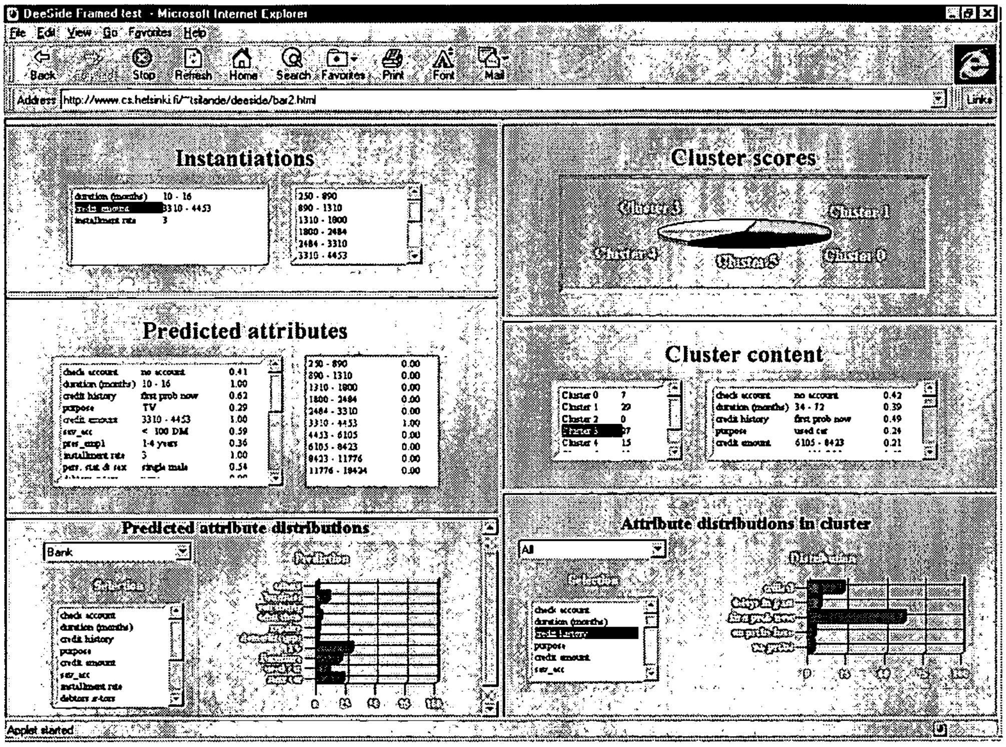


Figure 1: The classification performance of MDN and LD as a function of the sample size for the Effectiveness data set.

with another split for the training sample. Figure 1 gives the average (over 10 repetitions) performance of Mixture Density Networks and the linear discriminant methods as a function of the training sample size. Here we can see that both of the methods show quite similar small sample performance, the shapes of the curves being almost identical. However, this type of the analysis tells us about the complexity of modeling the data set with the model classes given. We can see that both methods are asymptotically approaching success rate of 40-50%, which is not particularly high.

6 MDN in exploratory analysis

In standard discriminant analysis, once the canonical discriminant functions have been derived, one can try to interpret their meaning. This is typically done by examining the relative positions of the data cases and group centroids, and by studying the relationships between the individual variables and the functions.



DeeSide Framed test - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back
Stop
Refresh
Home
Search
Favorites
Print
Font
Mail
Address http://www.cs.helsinki.fi/~tsilande/deeside/bar2.html
Links
Instantiations
Cluster scores
Cluster 3
Cluster 1
Cluster 4
Cluster 5
Cluster 0
Predicted attributes
Cluster content
Predicted attribute distributions
Bank
Attribute distributions in cluster
All
Applet started

duration (months)	10 - 16
credit amount	3310 - 4453
installment rate	3

250 - 890
890 - 1310
1310 - 1800
1800 - 2484
2484 - 3310
3310 - 4453

check account	no account	0.41
duration (months)	10 - 16	1.00
credit history	first prob now	0.62
purpose	TV	0.29
credit amount	3310 - 4453	1.00
sav_acc	< 100 DM	0.59
pres_empl	1-4 years	0.36
installment rate	3	1.00
pers. stat & sex	single male	0.54

250 - 890	0.00
890 - 1310	0.00
1310 - 1800	0.00
1800 - 2484	0.00
2484 - 3310	0.00
3310 - 4453	1.00
4453 - 6105	0.00
6105 - 8423	0.00
8423 - 11776	0.00
11776 - 18424	0.00

Cluster 0	7
Cluster 1	29
Cluster 2	0
Cluster 3	37
Cluster 4	15

check account	no account	0.42
duration (months)	34 - 72	0.39
credit history	first prob now	0.49
purpose	used car	0.24
credit amount	6105 - 8423	0.21

check account
duration (months)
credit history
purpose
credit amount
sav_acc
installment rate
debtors g-tors

check account
duration (months)
credit history
purpose
credit amount
sav_acc

Figure 2: A snapshot of the interface of the NONE software tool.

One of the methods is to study the structure coefficients of the variables to see how much a variable X_i has in common with a discriminant function f_{ld}. In our MDN approach the corresponding notion would be the Kullback-Leibler distance of the unconditional and conditional marginal likelihood of X_i, i.e.,

$$\mathcal{D}_{\mathrm{KL}}(p(X_i|X_m = k, \Theta), p(X_i|\Theta)), \text{where}$$

where $\mathcal{D}_{\mathrm{KL}}(p, q)$ is the relative entropy between p and q (Cover and Thomas, 1991). Similarly the corresponding notion to Wilk's lambda is the relative entropy between the unconditional and conditional joint distributions, i.e.,

$$\mathcal{D}_{\mathrm{KL}}(p(\vec{X}|X_m = k, \Theta), p(\vec{X}|\Theta)).$$

MDN networks model the joint probability distribution of the variables $X_1, \ldots, X_m$. Once we have built our model Θ, we can in fact *explore the* *predictive (marginal) distribution of any variable X_i* given the values of other variables, not just the group variable X_m. Modeling the full joint distribution gives us an extremely powerful exploratory tool— here we only want to briefly address some of the questions that can be answered by such a tool:

- **Variable predictive distributions for a given group.** In the extreme case we can fix in the data vector only the value of the group variable X_m, after which the MDN can calculate all the marginal predictive distributions. This means that one can study the distribution of any variable conditioned by the fact that the data vector $\vec{d}$ belongs to the group. For example in our Effectiveness data set we can fix one teacher education department value, and then explore what is the predicted attitude towards readiness for multimedia teaching for teachers that graduated from that particular department.

- **Variable predictive distribution of the group variable given some combination of other variable values.** We can reverse the situation in the previous item, and explore the effect of some value combination of variables $X_i, X_j, \ldots$ for predicting the group. Again, to give an example, we could explore which of the teacher education departments seems to have given the least readiness to teachers for using computers and multimedia in their teaching.

- **Variable predictive distribution of a non-group variable given some combination of other variable values.** Similarly, based on the implicit clustering induced by the group variable, one could also explore the predictive distribution of any non-group variable X_i given the values of some other non-group variables X_j, X_k, etc., without fixing the group value.

The MDN based approach has been implemented and runs on a Pentium PC under Linux operating system. Figure 2 illustrates the experimental software tool called NONE, which provides a flexible graphical interface for building MDN models, and exploring the predictive distributions. NONE is programmed in Java, and thus can be used with any Java compatible Internet browser. A running Java$^{\text{TM}}$ demo of the software can be accessed through our WWW homepage at URL "http: //www.cs.Helsinki.FI/research/cosco/".

7 Conclusion

In this paper we have discussed some of the methodological issues of using a class of neural networks, called Mixture Density Networks, for discriminant analysis. We demonstrated that, as opposed to many other neural network models, Mixture Density Networks have the advantage of having a rigorous probabilistic interpretation, and thus the resulting models can also be used for explorative purposes. In addition MDN have proven to be a viable alternative as a classification procedure in discrete domains, which is supported by the results in the empirical part of our work. The use of full joint probability models in discriminant analysis raises interesting methodological questions, some of which were addressed in our discussion. This paper has discussed ongoing research, and more extensive theoretical and experimental treatment e.g., in the context of factor analysis is a topic for future work.

Acknowledgements

This research has been supported by the Technology Development Center (TEKES), and by the Academy of Finland.

References

Baestaens, D., Van den Bergh, W., and Wood, D. (1994). *Neural Network Solutions for Trading in Financial Markets.* Financial Times / Pitman Publishing.

Bernardo, J. and Smith, A. (1994). *Bayesian theory.* John Wiley.

Bezdek, J. (1994). What is computational intelligence? In Zurada, J., II, R. M., and Robinson, C., editors, *Computational Intelligence - Imitating Life.* IEEE Press.

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Bishop, C. M. (1994). Mixture density networks. Technical Report NCGR/4288, Neural Computing Research Group, Department of Computer Science, Aston University.

Cheng, B. and Titterington, D. (1994). Neural networks: a review from a statistical perspective (with discussion). *Statist. Sci.*, 9:2–54.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* John Wiley & Sons, New York, NY.

Everitt, B. and Hand, D. (1981). *Finite Mixture Distributions.* Chapman and Hall, London.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors (1996). *Advances in Knowledge Discovery and Data Mining.* MIT Press, Cambridge, MA.

Gallant, S., editor (1993). *Neural Network Learning and Expert Systems.* MIT Press, Cambridge, Massachusetts.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation.* IEEE Press/Macmillan College Publishing Company, New York.

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

Hinton, G. (1989). Connectionist learning procedures. *Artificial Intelligence*, 40(1–3).

Hinton, G. (1992). How neural networks learn from experience. *Scientific American*, 267(104–109).

Hinton, G. and Sejnowski, T. (1983). Optimal perceptual inference. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 448–453, Washington DC. IEEE, New York, NY.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214.

Klecka, W. (1981). *Discriminant analysis.* Sage Publications, Beverly Hills, CA.

Kohonen, T. (1995). *Self-Organizing Maps.* Springer-Verlag, Berlin.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H., and Grünwald, P. (1997). Comparing predictive inference methods for discrete domains. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, pages 311–318, Ft. Lauderdale, Florida.

Mackay, D. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4:448–472.

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

MacKay, D. J. C. (1992b). The evidence framework applied to classification networks. *Neural Computation*, 4(5):698–714.

McLachlan, G., editor (1992). *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley & Sons, New York.

Niemi, H. and Tirri, K. (1996). Effectiveness of teacher education. new challenges and approaches to evaluation. Technical Report A 6/1996, Department of Teacher Education in Tampere University.

Niemi, H. and Tirri, K. (1997). Readiness for teaching profession evaluated by teachers and teacher educators. In Press.

Norušis (1990). *SPSS Advanced Statistics User's Guide.* SPSS Inc, Chigago.

Ripley, B. (1993). Statistical aspects of neural networks. In O.E. Barndorff-Nielsen, J. J. a. W. K., editor, *Networks and Chaos - Statistical and Probabilistic Aspects*, pages 40–123. Chapman and Hall, London.

Ripley, B. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B)*, 36:111–147.

Tirri, H., Kontkanen, P., and Myllymäki, P. (1996). Probabilistic instance-based learning. In Saitta, L., editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 507–515. Morgan Kaufmann Publishers.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Using Neural Networks for Descriptive Statistical Analysis of Educational Data

Author(s): Henry Tirri, Tomi Silander, Kirsi Tirri

Corporate Source: Univ. of Helsinki

Publication Date: 3/24/97

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ Sample sticker to be affixed to document

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document □

or here

Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature:

Printed Name:

Henry Tirri
Sr Research Scientist

Addre

Postal address: P.O. Box 26, Department of Computer Science
FIN-00014 University of Helsinki, FINLAND

Sreet address: Teollisuuskatu 23, 00051 Helsinki, Finland

Telephone: +358 9 708 44173 (Office)
+358 40 5000 533 (Mobile)

Telefax: +358 9 708 44441 (Department)
+358 9 708 44213 (CoSCo)

Email: Henry.Tirri@cs.Helsinki.FI
cosco@cs.Helsinki.FI

URL: http://www.cs.Helsinki.FI/~tirri/
http://www.cs.Helsinki.FI/research/cosco/

□

□

□

UNIVERSITY OF HELSINKI
Department of Computer Science
Complex Systems Computation Group (CoSCo)